on and similar papers at core.ac.uk

brou

provided by Elsev

Multiscale Analysis and Data Networks

A. C. Gilbert

AT&T Labs-Research, 180 Park Avenue, Florham Park, New Jersey 07901
E-mail: agilbert@research.att.com

Communicated by Henrique S. Malvar

The empirical finding of self-similarity in data network traffic over many time scales motivates the need for analysis tools that are particularly well adapted for identifying structures in network traffic. These structures span a range of time scales or are scale-dependent. Wavelet-based scaling analysis methods are especially successful, both collecting summary statistics from scale to scale and probing the local structure of packet traces. They include both spectral density estimation to identify large time-scale features and multifractal estimation for small time-scale bursts. While these methods are primarily statistical in nature, we may also adapt them to visualize the "burstiness" or the instantaneous scaling features of network traffic. This expository paper discusses the theoretical and implementation issues of wavelet-based scaling analysis for network traffic. Because data network traffic research does not consist solely of *analysis*, we show how these wavelet-based methods may be used to monitor and infer network properties (in conjunction with on-line algorithms and careful network experimentation). More importantly, we address what types of networking questions we can and cannot investigate with such tools. © 2001 Academic Press

1. INTRODUCTION

Scaling behavior appears to be a ubiquitous and inherent feature of data network traffic [9, 13]. It holds over a wide range of time scales and is present in a wide range of networks, from local area networks to wide area networks. Furthermore, while small aspects of this behavior change with the advent of new "killer apps," the gross features remain the same despite great changes in user behavior on these networks [10]. The description, analysis, and finally understanding of this scaling behavior is important for building accurate, parsimonious, and physically relevant models of data network traffic. The accuracy of traffic models is crucial for assessing network performance. There may be some performance metrics for which this scaling behavior produces unexpected results and there may be some which are not affected at all by these features.

Because wavelets and time-scale analysis more generally are well suited for scaling analysis, they are natural tools with which to describe and to probe network traffic. Such tools have flourished in this area recently and have yielded some very interesting

1063-5203/01 $35.00
Copyright © 2001 by Academic Press
All rights of reproduction in any form reserved.

results [2, 9]. However, to move beyond model fitting and parameter estimation for network traffic, we must begin to develop physical understanding of this scaling behavior. We must interpret what scaling behavior means for data networks.

The purpose of this paper is not to introduce new technical results or to prove theorems but to explain to the *applied and computational harmonic analysis* community how tools and concepts they are familiar with are used in a setting with which they may not be so familiar. There are many papers that focus on the technical and quantitative aspects of some of these methods ([2, 17] are but two examples) so this paper aims to show how these types of tools yield insight about data networks using the more qualitative aspects, including visualization. There are, of course, limitations to these tools. Some of these limitations arise from the finiteness of our datasets and some limitations are more fundamental. For example, we use these tools to study one-dimensional time series (or samples of a one-dimensional function) and we do not gather a complete picture of a network from such a simple set of observations.

We begin by pinning down several rigorous definitions of scaling behavior. Then we discuss the wavelet-based tools, giving precise mathematical results about these tools where possible and also highlighting the use of heuristics along with these rigorous theorems. We move to more empirical work showing what networking inferences we can make with these tools. Finally, we end with a discussion of future directions.

2. WHAT DOES SCALING MEAN?

To say that a time series or a process or a measure displays *scaling properties* means many different things depending on the context and the definition of scaling properties. This is especially true when we refer to scaling properties in networks and network traffic. We can talk about the power law scaling inherent in file sizes on servers [11, 16] or in the duration of user sessions [5] but we can also mean the scaling behavior of the traffic rate on a link (measured in the number of packets per time unit passing an observation point). Underlying these different definitions is the intuitive notion that the object we are studying (whether it is a process, measure, or function) has no inherent characteristic scale; it enjoys scale invariance.

Mathematically, the following scaling definitions depend critically on the type of object we are studying, be it a (random) measure, a time series, a function, or a (random) process. While this philosophical convention makes little difference in practice (data are simply samples of whatever underlying object we take as convention), it does matter when we try to use mathematical properties of our tools to deduce information about our data. In the following sections we outline the different scaling definitions used in analyzing network data traffic.

2.1. Self-Similar and Long Range Dependent Processes

DEFINITION 2.1. A process $\{X(t) \mid t \in \mathbf{R}\}$ is *self-similar with parameter* $H > 0$ if $X(0) = 0$ and $\{X(at) \mid t \in \mathbf{R}\}$ and $\{a^H X(t) \mid t \in \mathbf{R}\}$ have the same (finite-dimensional) distributions.

This definition excludes stationary processes but does allow for processes with stationary increments; that is, the finite-dimensional distributions of $\{X(t+s) - X(s) \mid t \in \mathbf{R}\}$ do not depend on s.

We will restrict ourselves to finite variance processes so that we may define the covariance and spectral density of self-similar processes. In particular, we wish to connect self-similar processes to long-range dependent processes and use a wavelet-based scaling analysis to analyze both types of processes. Let us assume that X is a self-similar process with independent increments and with a self-similarity parameter $0 < H < 1$. The variance of X is $\mathbf{E}(X^2(t)) = \sigma^2|t|^{2H}$ and the covariance is

$$\mathbf{E}\big(X(t)X(s)\big) = \frac{\sigma^2}{2}\big(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\big).$$

The increments of X, $Y(l) = X(l+1) - X(l)$, for $l = 0, 1, \ldots$, are stationary with mean zero and autocovariance

$$r(k) = \mathbf{E}\big(Y(l)Y(l+k)\big) = \frac{\sigma^2}{2}\big(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\big).$$

Note that for $H = 1/2$, $r(k) = 0$ for $k \neq 0$ and for $H \neq 1/2$, $r(k) \sim \sigma^2(H)(2H-1)k^{2H-2}$ as k tends to ∞.

Let $\Gamma(\lambda)$, $-\pi \leq \lambda \leq \pi$, denote the spectral density of Y. Then we can express Γ in terms of r using the relation

$$\Gamma(\lambda) = \frac{1}{\pi}\sum_{k=-\infty}^{\infty} r(k)e^{-ik\lambda}.$$

Observe that $r(k)$ tends to zero as $k \to \infty$ for all $0 < H < 1$ but for $1/2 < H < 1$, $r(k)$ decays so slowly that $\sum_k r(k)$ diverges. This affects the behavior of the spectral density Γ: if $0 < H \leq 1/2$, $\Gamma(0) = 0$ and if $1/2 < H < 1$, then $\Gamma(0) = \infty$. In particular, $\Gamma(\lambda)$ obeys a power law as $\lambda \to 0$ for $H \neq 1/2$,

$$\Gamma(\lambda) \sim \sigma^2 C(H)|\lambda|^{1-2H},$$

where $C(H)$ is a constant which depends only on H.

The typical example of such a self-similar process is *fractional Brownian motion (fBM)* $\{B_H(t) \mid t \in \mathbf{R}\}$ with $0 < H < 1$. This is a Gaussian self-similar process with finite variance. Its increments $Y_H(l) = B_H(l+1) - B_H(l)$, $l = 0, 1, \ldots$, are stationary and are called *fractional Gaussian noise (fGN)*. When $H = 1/2$, fBM and fGN are simply Brownian motion and Gaussian white noise, respectively.

Self-similar processes are intimately tied to long range dependent (LRD) processes or time series since they are one method for generating such LRD processes. LRD is defined in terms of second-order properties of a time series, so we must restrict our attention to finite variance time series.

DEFINITION 2.2. A finite variance second order stationary time series $\{Z(k) \mid k \in \mathbf{Z}\}$ has *long-range dependence (LRD)* if its spectral density $\Gamma(\lambda)$ obeys a power law scaling at low frequencies; i.e.,

$$\Gamma(\lambda) \sim C|\lambda|^{-\alpha} \qquad \text{as } \lambda \to 0$$

for C a constant and $0 < \alpha < 1$.

One can check that for X a self-similar process with $1/2 < H < 1$, the increments of X have LRD or asymptotic self-similarity with $\alpha = 2H - 1$. However, there are LRD time series that do not arise from self-similar processes (see [18] for an FARIMA$(0, d, 0)$ example).

2.2. *Fractals and Multifractals*

Loosely speaking, we refer to self-similarity as a global property of a process or (even more loosely) of a data set. The self-similarity parameter measures how the entire process scales from one time scale to another. We might also be interested in how a process (or its sample paths), a measure, or a function scale locally. There are two ways to describe this local behavior. The first is a summary of the local behavior (defined for measures) and the second is the local regularity of a function at a point.

We first define the local scaling exponent of a measure to capture its local regularity:

DEFINITION 2.3. Let μ be a measure on $\mathbf{R}$ and $supp(\mu)$ its support. The *singularity exponent* at $x_0 \in supp(\mu)$ is the limit (if it exists)

$$\alpha(x_0) = \lim_{\epsilon \to 0^+} \frac{\log \mu(B(x_0, \epsilon))}{\log \epsilon}, \tag{1}$$

where $B(x_0, \epsilon)$ is the ball of size ϵ centered at x_0.

While it is possible to calculate the local scaling exponents $\alpha(x_0)$ of a measure μ for each point $x_0 \in supp(\mu)$, it is not always the best way to examine the local properties of the measure; this is simply too detailed a perspective. Instead, we can develop a more refined approach from which we may draw statistically sound conclusions about how frequently certain scaling exponents appear. We refer to the statistical distribution of scaling exponents $\alpha(x_0)$ and the frequency with which $\alpha(x_0)$ takes on a specified value α as the multifractal spectrum of a measure μ. In general, the points in $supp(\mu)$ with equal singularity strength form subsets K_α of $supp(\mu)$ that themselves have fractal (geometric) properties (hence the notion "multifractal"). In other words, the multifractal structure of a measure μ refers to the Hausdorff dimensions (d_h) of the sets where the measure μ has scaling exponent α:

$$K_\alpha = \left\{ t \,\middle|\, \lim_{\epsilon \to 0^+} \frac{\log \mu(B(x_0, \epsilon)}{\log \epsilon} = \alpha \right\}.$$

DEFINITION 2.4. The function $f(\alpha) = d_h(K_\alpha)$ is called the *multifractal spectrum of* μ.

If μ has a continuous positive density on $supp(\mu)$, then the dimension of K_α as a function of α is a spiked function which takes on the value 1 at $\alpha = 1$. The function $f(\alpha)$ for a monofractal measure will also be a spiked function of α. For further examples and details, see, for example, [3].

There is special kind of multifractal measures which play an important role in our wavelet analysis: random multiplicative measures. While we are hesitant to model network traffic as conservative cascades, the statistical tools with which we analyze traffic traces are rigorous ones for this class of random measures.

DEFINITION 2.5. A *multiplicative cascade* is an iterative process that fragments a given set into smaller and smaller pieces according to some geometric rule and, at the same time, distributes the total mass of the given set according to another rule (usually a random fraction of mass at the previous step). If the mass redistribution rules preserves the total mass of the initial set at each stage of the construction almost surely, we call this a conservative cascade. We call the random variable W that specifies the random fraction the generator of the conservative cascade.

Unfortunately, it is not possible to define multifractal processes in terms of the local behavior of sample paths of the processes. There are examples of self-similar processes (e.g., linear fractional stable motion) which have continuous paths for some range of the self-similarity parameter and for other choices of the parameter have paths that are unbounded on any finite interval. We can define multifractal processes in terms of the scaling properties of the moments of the process however (see [14] for more details).

DEFINITION 2.6. Let T and Q be intervals on $\mathbf{R}$ with positive lengths, $0 \in T$, and $[0, 1] \subset Q$. If $\{X(t) \mid t \geq 0\}$ has stationary increments and finite variance, $X(0) = 0$, and

$$\mathbf{E}(|X(t)|^q) = C(q)t^{\tau(q)+1} \qquad \text{for all } t \in T,\ q \in Q,$$

then we say that X is a *multifractal process*. Note that $C(q)$ and $\tau(q)$ are deterministic functions.

A self-similar process X satisfies this definition trivially since the moments of X are given by $\mathbf{E}(|X(t)|^q) = \mathbf{E}(|X(1)|^q)|t|^{qH}$. Here the functions $C(q) = \mathbf{E}(|X(1)|^q)$ and $\tau(q) = qH - 1$ are simple to calculate.

Unfortunately, we often have only one data set which we must treat as a sample path of some underlying process and the set of processes whose sample paths are amenable to local regularity study is too small for our needs. Thus we turn to the local regularity characterization of functions and use the convention that our data set consists of samples of a function rather than a process or a measure.

2.3. Local Scaling Exponents and Regularity

DEFINITION 2.7. Let f be a function $\mathbf{R} \to \mathbf{R}$, x_0 be a real number, and α be a strictly positive real number. We say f belongs to $C^\alpha(x_0)$ if we can find a polynomial P_m of degree $m < \alpha$ such that, for all x in a neighborhood of x_0,

$$|f(x) - P_m(x - x_0)| \leq C|x - x_0|^\alpha. \tag{2}$$

Then the *pointwise Hölder exponent* of the function f at x_0 is defined by

$$H(f, x_0) = \sup\{\alpha > 0 \mid f \in C^\alpha(x_0)\}. \tag{3}$$

Note that such a polynomial P_m can be found even if a Taylor development of f around x_0 does not exist.

The relation between the singularity exponent of a measure μ and the pointwise Hölder exponent is the following: if we denote by $F(x) = \int_0^x d\mu$, then the pointwise Hölder

exponent of F at each point x_0 is the singularity exponent of μ at x_0. Thus, according to this last remark, we will refer to the Hölder or singularity exponents of any measure interchangeably.

Intuitively, this exponent α is the exponent of a parabolic envelope around x_0 in which the graph of the considered function f lies and evolves locally. That does not imply anything more, and unfortunately that is not enough to describe completely the behavior of f around x_0. The best way to grasp how different two functions with the same Hölder exponent can be is to study the following well-known example. Let us define

$$f_1(x) = -|x|^3 \text{ if } x < 0, \qquad \text{and} \qquad |x|^2 \text{ if } x \geq 0, \tag{4}$$

$$f_2(x) = |x|^2 \sin(1/x^2), \qquad \text{and} \qquad f_2(0) = 0. \tag{5}$$

Both functions have a Hölder exponent $H(f_i, 0) = 2$ at 0, since one can write $|f_1(x) - x^2| \leq |x|^2$ and $|f_2(x)| \leq |x|^2$ in a neighborhood of 0. But while f_1 has a continuous derivative, the derivative of f_2 explodes as x gets closer to 0.

2.4. Cusps versus Chirps

Arneodo *et al.* [4] build a framework to categorize and to describe these different behaviors. (This framework is formalized and extended significantly in [15].) The main idea is the following: if a function f around a point x_0 has an Hölder exponent α, and if it is "regular" enough, then the derivative of f will have an Hölder exponent $\alpha - 1$, or, almost equivalently (but more stable from a numerical point of view), the integral of f will have an Hölder exponent $\alpha + 1$. This corresponds to the intuitive idea of smoothness; i.e., we lose one degree of smoothness when differentiating; we win one degree when integrating. For some functions we win more than one degree when integrating a function: the above function f_2 is a typical illustration of this fact. More generally, consider any function which belongs to the set of functions of the type $g_{h,\beta}(x) = |x - x_0|^h \sin(1/|x - x_0|^\beta)$, that oscillate infinitely fast around the point x_0. Then when integrating such a function, we obtain the function $F(x) = \int_{x_0}^{x} g_{h,\beta}(t)\, dt$ that has a local Hölder exponent at x_0 bigger than expected (exactly $h(1+\beta)$) because of cancellation effects. This leads to the following definition:

DEFINITION 2.8. Let f be a function $\mathbf{R} \to \mathbf{R}$. A *cusp* or *nonoscillating singularity* for f is a point x_0 where

$$H(f, x_0) = H\left(\int f, x_0\right) - 1. \tag{6}$$

A *chirp* or *oscillating singularity* for f is a point x_0 where

$$H(f, x_0) < H\left(\int f, x_0\right) - 1. \tag{7}$$

We can see that the Hölder exponent does not fully characterize the local behavior of a function. Thus, in order to complete the description, we need one more exponent, the *oscillation exponent* β, defined by $\beta = H(\int f, x_0) - 1 - H(f, x_0)$.

3. WAVELET-BASED SCALING ANALYSIS TOOLS

We present a very brief discussion of basic wavelet definitions, primarily to set our notation. We focus on compactly supported wavelets on **R**, which are particularly well adapted to sampled functions of finite length. We use the so-called *wavelet decomposition*; i.e., if ψ is the mother wavelet and f is the considered function,

$$f = \sum_{j,k} d_{j,k}\psi_{j,k}, \tag{8}$$

where

- $\psi_{j,k}$ is the translated and dilated version of ψ, namely $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$.
- $d_{j,k} = \int \psi_{j,k}(x) f(x)\, dx = \langle \psi_{j,k}, f \rangle$ is the *wavelet coefficient* of f at scale j at the point k.

The above equality is true under some mild assumptions on ψ; for details see [6]. Intuitively the amplitude of the wavelet coefficient $d_{j,k}$ measures the signal content around time $2^{-j}k$ at frequency $2^j * |support(\psi)|$. An important property of the wavelet is its number of vanishing moments. Indeed, if we assume that

$$\int x^i \psi(x)\, dx = 0, \qquad \text{for } i = 0, 1, \ldots, N-1, \tag{9}$$

then ψ is said to have N vanishing moments.

3.1. Self-Similarity (and LRD) and Energy Plots

Wavelets are a particularly good analysis tool for both self-similar and LRD processes because they are themselves scale invariant. Wavelets also turn the long range dependency inherent in these processes into short range stationary processes in the wavelet coefficients, yielding efficient spectral estimators. We first examine the statistical properties of the wavelet coefficients of self-similar processes. We shall see that these properties are very similar to those for LRD processes, thus tying these two types of processes closer together. We use these properties to define and to characterize the energy plots. In what follows we assume that the wavelet ψ is compactly supported (not a necessary condition for the mathematical results but a useful computational assumption) and that it has N vanishing moments. The following theorems are a collection of results from [1, 2, 7].

THEOREM 3.1. *Let X be a self-similar process. Then the wavelet coefficients $d_{j,k} = \langle \psi_{j,k}, X \rangle$ of X satisfy*

$$d_{j,\cdot} = 2^{-j(H+1/2)} d_{0,\cdot},$$

where equality holds in the distributional sense. At each fixed scale j, the wavelet coefficients $d_{j,k}$ form a mean zero short range dependent stationary process. The variance of the coefficents is given by

$$\mathbf{E}(d_{j,k}^2) = 2^{-j(2H+1)} C(H, \psi)\sigma^2 \qquad \text{for all } k$$

and the covariance structure by

$$\mathbf{E}(d_{j,k}d_{j',k'}) \sim |2^{-j}k - 2^{-j'}k'|^{2H-2N} \qquad \text{as } |2^{-j}k - 2^{-j'}k'| \to \infty.$$

THEOREM 3.2. *Let X be a long range dependent process with power law parameter α and assume that the analyzing wavelet has $N \geq \alpha/2$ vanishing moments. Then the wavelet coefficients $d_{j,k}$ of X are at each scale j a mean zero short range dependent process with variance*

$$\mathbf{E}(d_{j,k}^2) = \int \Gamma(\lambda)|\hat{\psi}(2^{-j}\lambda)|^2\, d\lambda \sim 2^{-j\alpha} C \int |\lambda|^{-\alpha}|\hat{\psi}(2^{-j}\lambda)|^2\, d\lambda \qquad \text{as } j \to \infty$$

and covariance

$$\mathbf{E}(d_{j,k}d_{j',k'}) \sim |2^{-j}k - 2^{-j'}k'|^{\alpha-1-2N} \qquad \text{as } |2^{-j}k - 2^{-j'}k'| \to \infty.$$

DEFINITION 3.1. To draw the energy plot, we first calculate the average energy at each scale j

$$E_j = \frac{1}{N_j} \sum_{k=0}^{N_j-1} |d_{j,k}|^2,$$

where N_j signifies the number of wavelet coefficients at scale j. Next we plot $\log(E_j)$ as a function of scale j.

The estimates E_j reveal the behavior of the second moment of the process X at each scale j. They are a non-parametric unbiased estimator for the variance of the wavelet coefficients $d_{j,k}$. In fact, the short range dependence structure of the wavelet coefficients implies that E_j is a near-optimal estimator for the second order behavior of X at scale j, with little coupling from the other scales. The power law structure of $\mathbf{E}(d_{j,k}^2)$ suggests that the energy plot for a self-similar or an LRD process is a straight line with slope $2H + 1$ or α. For the details of this estimator, see [2]. It is important to note that $\log \mathbf{E}(E_j)$ need not equal $\mathbf{E}(\log E_j)$ so care must be taken to estimate the scaling exponent of a process accurately.

While these mathematical results are careful precise statements, they often do not match what we see in practice. We frequently see energy plots with scaling over a range (but not all) of scales, two different scaling regions, and nonlinear graphs. While much statistical work can be done to accurately characterize and estimate such behavior, we focus on less precise but more network-oriented analysis. To motivate the type of network inferences we make in the following section, we start with an exactly self-similar process X (given in terms of the number of packets per time unit) and we modify it in a precise way so as to effect a particular departure from nonlinearity it the energy plot. We *replace* every τ samples with a constant value, mimicking a deterministic source embedded in the traffic that sends packets at a fixed rate. Figure 1 shows such a time series in the upper left plot. Simply by looking at this modified time series, we are unable to detect the presence of this periodic source (nor can we tell that the time series is self-similar). Similarly, if we look at the magnitude of the Fourier transform of this time series (upper right plot in Fig. 1), we can see the periodic component (highlighted large spikes in the spectrum) but we cannot

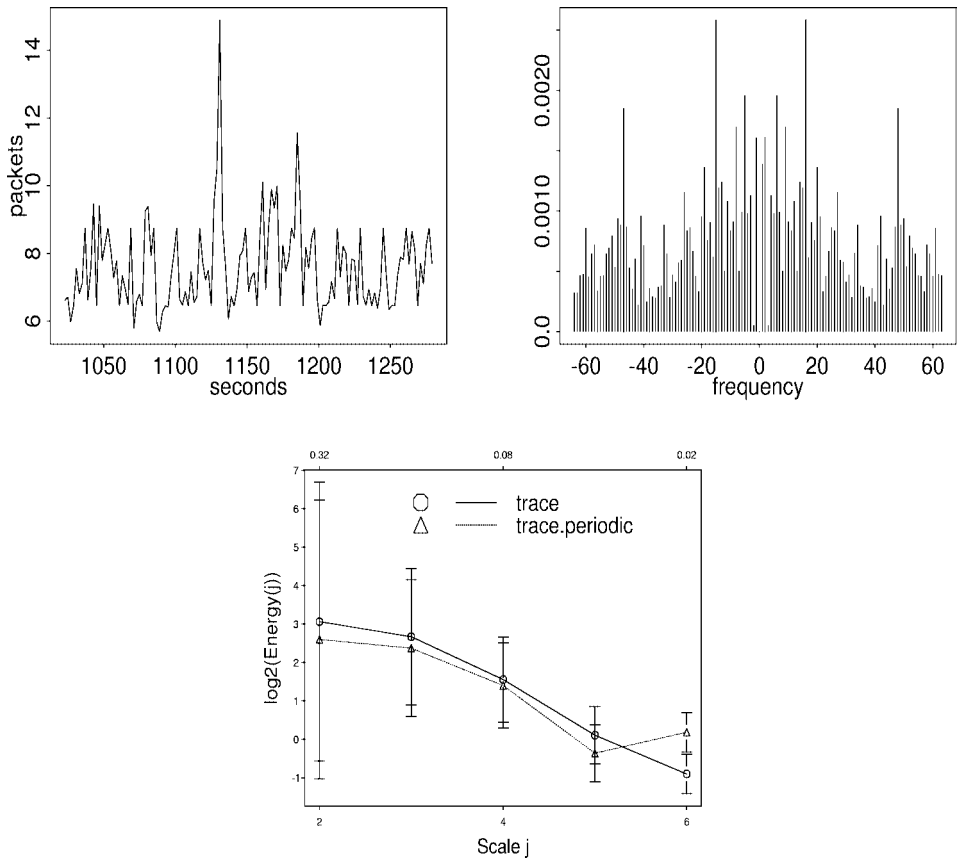

FIG. 1. A modified LRD time series (upper left); the magnitude of the Fourier transform of the time series (upper right); the energy plots of the original and modified time series (bottom).

tell that the rate process is self-similar. Only when we examine the energy function can we tell that the original time series is self-similar (curve labeled *trace* in the bottom plot of Fig. 1) because its graph is a straight line. We can also see that the modified time series (curve labeled *trace.periodic*) yields a linear energy function except at the dyadic scales surrounding the scale τ of the periodic component. Note that the periodic component is not simply added to the self-similar time series so we do not get an increase in energy at the scale τ, rather we get the more complicated behavior of a dip in the energy plot. If we introduce a periodic component at a larger scale, we see the dip in the energy curve move towards the left (larger scales). The curve does not appear to have two scaling regions; it simply looks like a straight line with an indentation at one scale.

3.2. Multifractals and Partition Functions

In Subsection 2.2 we defined the local scaling properties of a measure. Rather than compute the local scaling exponents at each point, we can collect summary statistics about the local behavior by computing the partition function. The wavelet-based partition

function is defined as

$$S(q, j) = \sum_k |C_j d_{j,k}|^q,$$

where C_j is a normalizing constant to remove the L^2 normalization of the wavelets. The partition function computes the qth order moments of the wavelet coefficients as a function of scale. We plot $\log S(q, j)$ versus j which gives us a family of curves, indexed by q. Next, we set the structure function $\tau(q)$ equal to the slope of $\log S(q, j)$ for each q.

An intuitive explanation for why the wavelet-based partition function should capture the local scaling behavior of a measure or a function relies upon the results in the first part of Subsection 3.3. Because a small local scaling exponent (i.e., a particularly abrupt burst) generates a large wavelet coefficient, its presence is magnified by taking qth powers across scales. Similarly, when a measure or function is locally smooth, its wavelet coefficients are small and diminished in the qth moments. We can make this analysis more precise.

THEOREM 3.3. *Let μ be a conservative cascade with generator W and $d_{j,k}$ its Haar wavelet coefficients. The variable $Z(q, j) = (\log S(q, j))/(-j \log 2)$ is an almost surely consistent estimator for $\tau(q) = -1 - \log_2 \mathbf{E}(W^q)$ as $j \to \infty$, provided $q < q^*$. The critical power q^* depends on the generator W (see* [17] *for details).*

In some cases, we can take the Legendre transform of $\tau(q)$ to compute $f(\alpha)$, the multifractal spectrum of the measure μ. However, in practice, the conditions necessary for such precision are impossible to verify so we use a more heuristic approach. We say that a fractal measure has a linear structure function (i.e., the slope of $\log S(q, j)$ changes linearly in q) while a multifractal has a nonlinear structure function. Such distinctions are frequently difficult to make with a high degree of confidence for realistic traffic traces so, as with the energy plots, we say that a nonlinear structure function denotes "interesting" local scaling, a departure from monofractal scaling. To illustrate these ideas, Fig. 2 shows the partition functions for an exactly self-similar time series (treated as sample values of a function, upper right) and for a conservative cascade (a multifractal measure with "interesting" local scaling, upper left). The lower plots depict the corresponding structure functions. We have added lines to the structure functions to illustrate how close or far they are from linear. Note that in the case of the conservative cascade, we have not simply fit a line to the structure function, we have extrapolated the behavior at high moments from those at lower moments. We then created families of curves on the partition function plots that would produce such extrapolated structure functions. Observe that for the exactly self-similar time series, the structure function is essentially a straight line and the extrapolated partition function curves lie very close to the actual curves unlike those for the conservative cascade.

3.3. Local Singularities and Blur Plots

The previous wavelet-based tools generate summary statistics and now we turn to wavelet-based local analysis. First we present the intuitive idea that explains why a wavelet analysis of a function gives information about its local singularities. Unfortunately, this intuition has reached the status of folklore and the rigorous technical results are a bit more complicated.

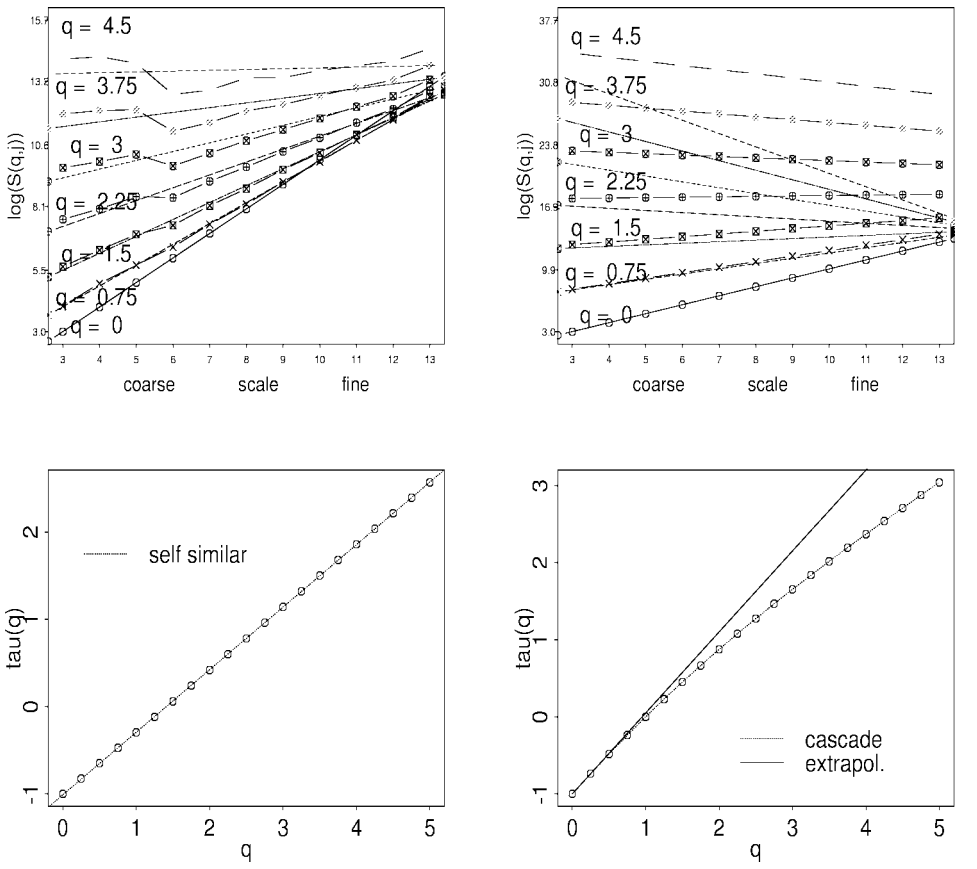

FIG. 2. The structure function (upper left) and partition function (lower left) of an exactly self-similar function; the structure function (upper right) and partition function (lower right) of a multifractal cascade.

If around a point x_0, f can be written as

$$f(x) = a_0 + a_1(x - x_0) + \cdots + a_{m-1}(x - x_0)^{m-1} + o((x - x_0)^m), \tag{10}$$

where $m < N$, then, for k such that $2^{-j}k$ is close to x_0,

$$d_{j,k} = \int 2^{j/2}\psi(2^j x - k) f(x)\, dx \tag{11}$$

$$= 2^{j/2} \int \psi(2^j x - k) o((x - x_0)^m)\, dx \tag{12}$$

$$= \mathcal{O}(2^{-j/2} 2^{-jm}). \tag{13}$$

This important result tells us that the decay of the wavelet coefficients around x_0 is closely related to the local regularity of the function f at this point, and also that the wavelet can catch the regularity only up to degree N. Note that from now on we will consider rescaled coefficients; i.e., $d_{j,k_{\text{new}}} = 2^{j/2} d_{j,k_{\text{old}}}$, to remove the normalization factor $2^{-j/2}$ in the previous estimate.

Let us recall the definitions of cusps and chirps so that we can understand why we claim that size estimates on wavelet coefficients of a function cannot *completely* characterize its local behavior. Indeed, on the one hand, if a function f admits a cusp h at x_0, then the equivalence $|d_{j,k}| \sim 2^{-jh}$ is optimal and one can prove that the maxima of the $|d_{j,k}|_{k\in\mathbf{Z}}$ lie in a "straight" cone pointing to x_0, defined by $|2^{-j}k - x_0| \leq \mathcal{O}(2^{-j})$. On the other hand, if f admits a chirp (h, β) at the point x_0, the wavelet coefficients will decrease as $2^{-jh(1+\beta)}$. In this case one can show that the maxima of the modulus of the wavelet coefficients lie in a "parabolic" cone that is wider than the "straight" cone and that is given by the equation $|2^{-j}k - x_0|^{1+\beta} \leq \mathcal{O}(2^{-j})$. For details, see [4].

The conclusion of these remarks is the following: in order to describe efficiently the local behavior of a function, we must take into account the amplitude of the wavelet coefficients $d_{j,k}$ associated to their localization in time (represented by k). The following fundamental theorem [12] confirms our discussion:

THEOREM 3.4. *Let f be a function: $\mathbf{R} \to \mathbf{R}$ and k_0 denote the index such that $2^{-j}k_0 = x_0$.*

- *If $f \in C^h(x_0)$, then*

$$|d_{j,k}| \leq C2^{-jh}(1 + |k - k_0|)^h. \tag{14}$$

- *If f satisfies* (14) *and if $f \in C^\epsilon(x_0)$, for any $\epsilon > 0$, then there exists a polynomial P_m of degree up to h, such that, if $|x - x_0| \leq 1/2$,*

$$|f(x) - P_m(x - x_0)| \leq C|x - x_0|^h \log(1/|x - x_0|). \tag{15}$$

Unfortunately this theorem does not provide a pure characterization, since the reverse implication introduces a logarithm factor. Meyer [15] obtains another result by imposing a global continuity condition, using the local two-microlocal spaces $C^{h,-h}(x_0)$. These spaces offer the advantage of having a wavelet-based characterization. Indeed, f locally belongs to $C^{h,-h}(x_0)$ if and only if $|d_{j,k}| \leq C2^{-jh}(1 + |k - k_0|))^h$ for all pairs (j, k) satisfying $|k - k_0| \leq 2^j$.

THEOREM 3.5. *Let f be a function: $\mathbf{R} \to \mathbf{R}$. Let ω be a positive continuous increasing function such that $\omega(y) = \mathcal{O}((\log 1/y)^{-N})$ if $|y| \leq 1/2$. Let us define $C_\omega(\mathbf{R}) = \{g \in C^0(\mathbf{R}) \,|\, |g(x + y) - g(x)| \leq \omega(|y|)$ for all x, $y\}$.*

- *If $f \in C^h(x_0)$, f locally belongs to $C^{h,-h}(x_0)$.*
- *Conversely, assume that $f \in C_\omega(\mathbf{R})$. Then if $f \in C^{h,-h}(x_0)$, $f \in C^{h-\epsilon}(x_0)$ for all $\epsilon > 0$.*

Now we can write the equality

$$H(f, x_0) = \sup\{h > 0 \,|\, f \in C^{h,-h}(x_0)\} \tag{16}$$

whenever $f \in C_\omega(\mathbf{R})$ (this is the global continuity condition we discussed before). We now have an effective characterization of the pointwise Hölder exponent by a size estimate on wavelet coefficients, using the inequality (14).

Using these theoretical results, we build a simple estimator for the pointwise Hölder exponent of a function, whether it is oscillating or not (for details see [19]). For network

data, we have a wide range of sampled functions of finite length and we want an estimate of local regularity for each sample point. We use a discrete nondecimated wavelet transform rather than an exact transform. While the exact transform completely characterizes the local behavior, for sampled functions of finite length, the oversampled or redundant discrete transform provides more information about the function with which to form a more stable estimator.

Let us assume that our signal consists of 2^n sample values. We denote by $d'_{j,k}$ the nondecimated wavelet coefficients of the signal. The inequality (14) in that case slightly changes and becomes

$$|d'_{j,k}| \leq C(2^{-j} + |k - k_0|/2^n)^{\alpha}. \tag{17}$$

To capture the pointwise Hölder exponent, we shall not focus solely on the maxima of the wavelet coefficients at each scale. The maxima are sufficient when analyzing a cusp singularity; however, they are not when dealing with oscillating singularities.

For a nonoscillating singularity it is sufficient to take the maximum only over the coefficients $d'_{J,k}$ at scale J to find the most important coefficient in inequality (17), while for an oscillating singularity the decay rate of the wavelet coefficients necessitates a more subtle analysis. In other words, we may use for a nonoscillating singularity a method motivated by Theorem 3.4 (a modulus-maxima approach) while this method may fail to produce an accurate estimate for an oscillating singularity. To analyze real data sets, we may consider noise that appears as oscillatory. For network traffic traces that are not noisy but rather extremely bursty, a more careful estimation than a simple modulus-maxima approach in necessary. We aim to extract the singularity exponent α from an estimate on the size of the wavelet coefficients. A simple way to achieve our goal is the following algorithm:

Construction. Let f be a function: $[a, b] \to \mathbf{R}$ and assume that $f \in C_{\omega}(\mathbf{R})$. Let $d'_{j,k}$ be its nondecimated (i.e., redundant) wavelet transform.

- Plot on the same picture, for each $j > 0$, the parametric curve (the parameter is k):

$$x_j(k) = \log_2(2^{-j} + |k - k_0|/2^n)$$
$$y_j(k) = \log_2(|d'_{j,k}|).$$

- Find all straight lines $\mathcal{D}$: $y = \alpha x + C$ such that:

  (1) $\mathcal{D}$ is above all the plotted points; i.e., $\forall j, \forall k$, $y_j(k) \leq \alpha x_j(k) + C$, and

  (2) $\mathcal{D}$ "touches" one of the parametric curves; i.e., there exists a sequence of pairs (j_m, k_m) such that $\lim_{m \to +\infty} y_{j_m}(k_m) - (\alpha x_{j_m}(k_m) + C) = 0$.

- Consider α_{max}, the maximum of the slopes α over all the straight lines $\mathcal{D}$ satisfying Properties (1) and (2).

4. NETWORK INFERENCES

As stated in the Introduction, the purpose of this paper is to demonstrate the abilities of wavelet-based scaling analysis tools for network inferences. Feldmann *et al.* [8] explore the role of variability in network operations and we cover some of the same experiments

in this section to convey the power and the pitfalls associated with these tools. Using a set of detailed experiments combined with the analysis and interpretation of real traffic traces, we show that using the energy function plots, we can infer the predominant user session characteristics and the predominant round trip time in the network. With the partition function plots, we can visualize the difference between closed and open feedback loops in two common data transfer protocols (TCP and UDP). We use several network configurations with a cloud of clients and servers at opposite ends of a bottleneck link. We are able to vary link attributes such as delay, bandwidth, and access speed. We can also modify the protocol features used to transfer files from the servers to the clients. Finally, we choose different distributions for the web session attributes such as the number of pages requested during a session and the number of objects per page. At the bottleneck link we measure the traffic rate process and examine the scaling properties of this process, given that we know exactly the network and user configurations.

4.1. User Profile

We know from Subsection 3.1 that if a time series obeys self-similar scaling or long range dependence, then the energy function graph will be a straight line (over those time scales which show scaling properties) with positive slope. In addition, if a time series obeys short range dependence, its energy function plot will show a horizontal line. In the first configuration of our experiments, the user session attributes (e.g., number of pages per session) are all distributed as exponential and in the second as Pareto distributions (with finite mean but infinite variance). We expect that the first configuration will generate traffic that is consistent with short range dependence while the second configuration will show self-similar scaling. Briefly, the authors of [13] show that a superposition of heavy-tailed ON/OFF sources generates a self-similar rate process. Figure 3 confirms our intuition.

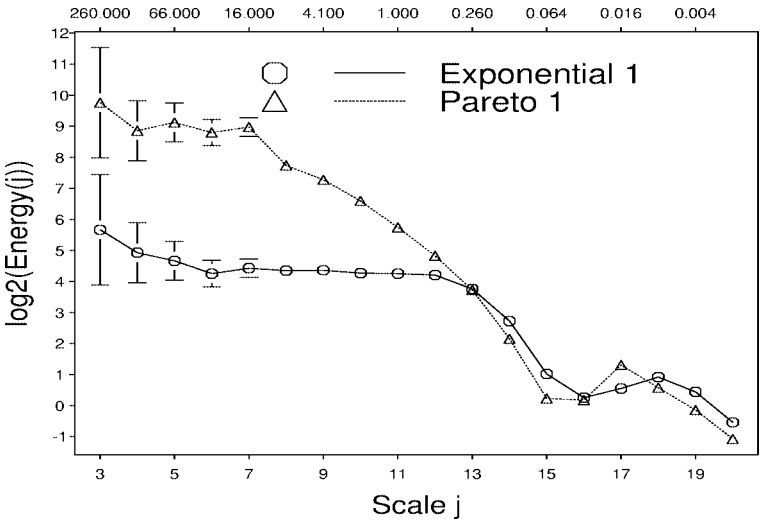

FIG. 3. The energy plot of the traffic rate process from infinite variance user profiles.

4.2. *Network Time Issues*

While the two types of user session distribution explain the differences in the slopes at large time scales in Fig. 3, they do not explain the behavior in the graphs at small time scales. At these time scales we see a predominant dip in both energy function plots. This dip occurs around the 16- to 32-ms time scale. Our previous mathematical discussion suggests that these time series contain a strong periodic component. Our physical intuition points to the predominant round trip time (RTT) in the network as the mechanism generating the periodic component. Because the experimental configuration is a homogeneous one where all sources experience roughly the same RTT, we expect to see a periodic component at that RTT and a dip in the energy plot at the smallest dyadic scale that is at least as large as the RTT. To verify our intuition, we perform this experiment with two network configurations: one with an RTT of 24 ms and the second with an RTT of 1.3 s. We effect these RTTs by changing the bottleneck link delay from 1 to 640 ms. The energy plots for these two configurations are shown in Fig. 4. Note the large dips in the two plots corresponding to the two RTTs. If we perform a similar experiment and use a heterogeneous network environment instead with a range of server link delays, we get a range of RTTs resulting in a wider dip in the energy function plot, centered about the typical RTT in the network. We see a similar picture if the network is heavily congested and packets experience a variety of delays. Thus the energy function plots reveal two types of network behavior: infinite variance user sessions (via lines with nonzero slope) and prominent RTTs (through nonlinear features, especially at small time scales).

4.3. *Network Protocol and Complexity*

While the energy function plots can tell us about the user profile and network timing issues, they cannot tell us everything. We turn to the partition function to infer protocol characteristics, such as closed versus open loop congestion control, and to discern some

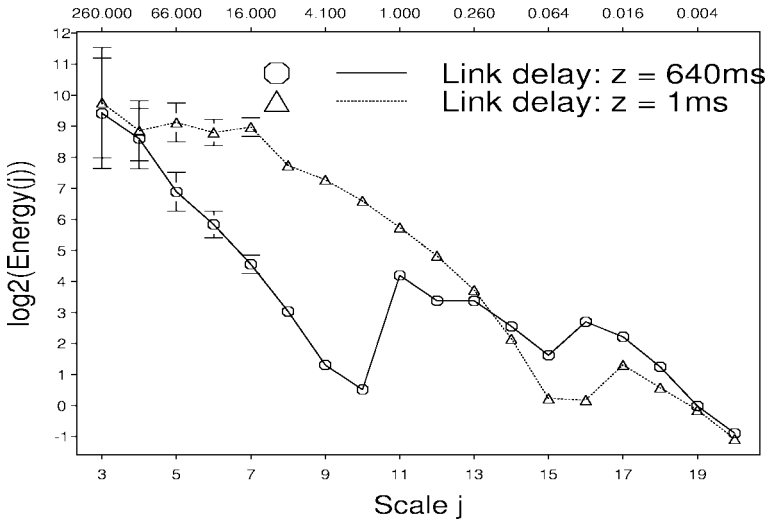

FIG. 4. Energy function plot of rate process from network configurations with two different predominant RTTs.

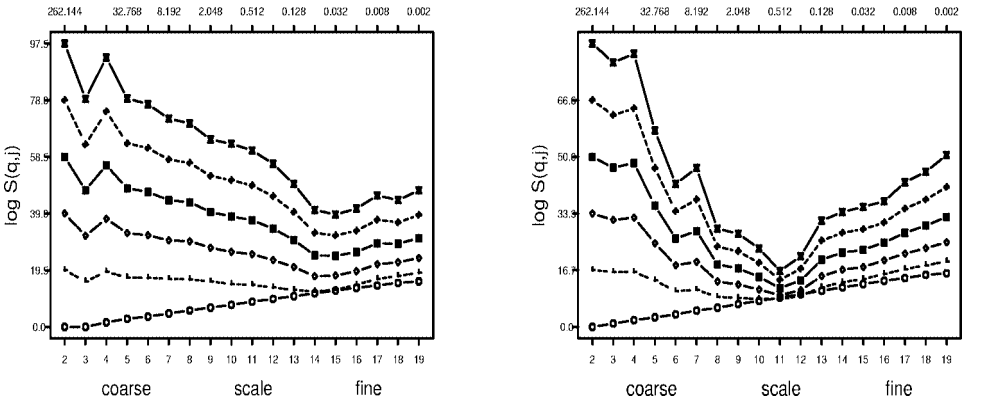

FIG. 5. Partition function plots for sources transmitting packets via UDP (left) compared with TCP (right).

degree of network topological complexity. Recall from our previous discussion that we are looking for a nonlinear relationship between the slopes of the curves in the partition plot and q the power to which we raise each wavelet coefficient $d_{j,k}$.

We perform two experiments: one in which the sources use UDP for transmitting packets to the receivers and the other in which they use TCP. The main difference between these two protocols is that UDP is an unreliable protocol; sources simply send packets at a constant rate regardless of congestion or loss of those packets. With TCP, a source receives acknowledgements from the receiver indicating that packets have indeed reached their destination. In addition, sources which transmit via TCP resend any lost packets, providing a reliable data transfer protocol, and they throttle their sending rate in the face of losses. We say that TCP has a closed loop congestion control while UDP has an open loop. Figure 5 shows the partition functions for sources using UDP (left) and those using TCP (right). For the UDP sources, one can see some dipping in the family of curves because of the way UDP sources periodically inject packets into the network. One can also check that the slopes of these curves for the UDP sources is proportional to q while the slopes of the curves (measured over fine scales) for the TCP sources are nonlinear in q. We can use this type of information to detect the fingerprints of predominant network applications. Streaming applications such as RealAudio often use UDP while HTTP sits on top of TCP. Also we can use this type of inference to detect misconfigured (or misbehaving) TCP sources. There is also some evidence to suggest that the more complicated the network topology (for example, two-way traffic on a well-connected graph), the more the partition function shows a nonlinear relationship between the slopes and q (see [8]).

4.4. Local Irregularity

We end with two time series representing the number of packets per 10 ms from a local area network (LAN) and a wide area network (WAN) packet trace. While we are not yet able to make precise network inferences from such local regularity analysis, the plots generated are informative and suggest future work. The time series and their Hölder exponent estimates are show in Fig. 6. We find that the LAN exponent estimates are more tightly clustered around their mean value of approximately 0.8 than the WAN estimates. This finding is compatible with the observation that LAN network traffic is consistent

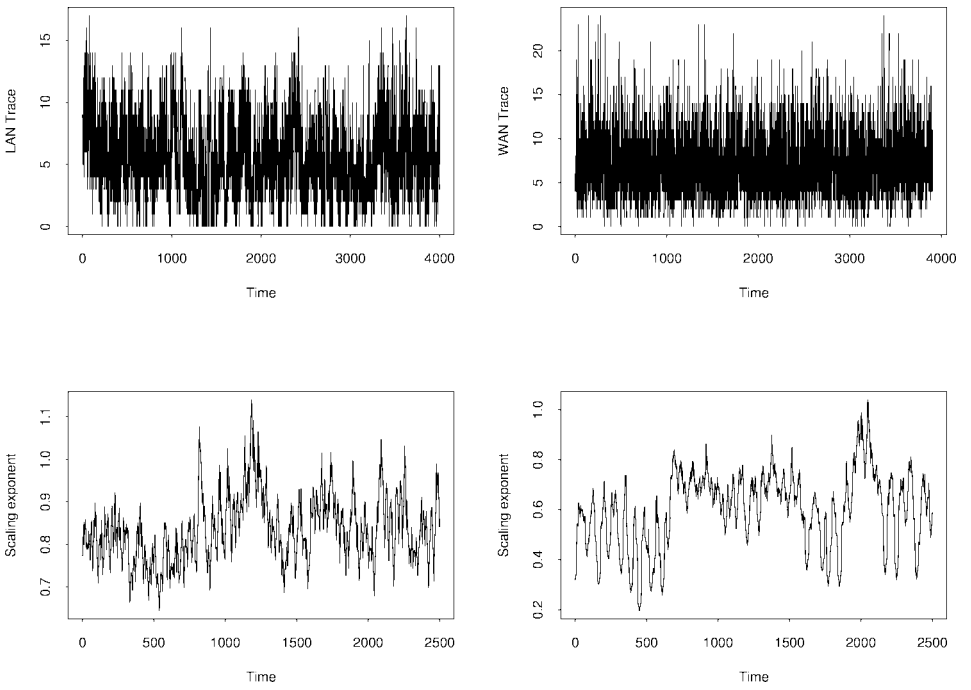

FIG. 6. Top left, LAN trace (number of packets per time unit). Top right, WAN trace (number of packets per time unit). Bottom left, computed scaling exponents for the LAN trace. Bottom right, computed scaling exponents for the WAN trace.

with self-similar processes [13]. In addition, the WAN Hölder exponents take on smaller values than the LAN exponents, indicating more burstiness in the WAN than the LAN traffic. Also, there are time periods during which the WAN Hölder exponents are tightly clustered and some for which the exponents are fairly evenly distributed over the range $0.2 \leq \alpha \leq 0.8$.

5. CONCLUSION

These wavelet-based scaling analysis tools are incredibly useful for describing and detecting certain kinds of properties of one-dimensional functions, measures, or random processes. We can compute summary statistics about scale-dependent properties, local scaling behavior, and even extremely localized information about the local regularity of network traffic. Because these methods can be implemented in an on-line fashion, we can use them to monitor network links, either at one main link or at many access points.

However, these tools do have some serious drawbacks when it comes to the next step in network measurements. They work only with information local to a network. They cannot address distributed network measurements at all. We cannot take measurements of any granularity from all over our network and derive some spatio-temporal scaling picture. We cannot even incorporate any knowledge of network topology into these local measurements.

REFERENCES

1. P. Abry, P. Gonçalvès, and P. Flandrin, Wavelets, spectrum analysis and $1/f$ processes, *in* "Wavelets and Statistics" (A. Antoniadis and G. Oppenheim, Eds.), Lecture Notes in Statistics, Vol. 105, pp. 15–30, Springer-Verlag, New York/Berlin, 1995.
2. P. Abry and D. Veitch, Wavelet analysis of long-range dependent traffic, *IEEE Trans. Inform. Theory* **44** (1998), 2–15.
3. A. Arneodo, "Wavelets: Theory and Application," pp. 349–502, Oxford Univ. Press, New York, 1996.
4. A. Arneodo, E. Bacry, S. Jaffard, and J. F. Muzy, Singularity spectrum of multifractal functions involving oscillating singularities, *J. Fourier Anal. Appl.* **4**, No. 2 (1998), 159–174.
5. M. E. Crovella and A. Bestavros, Self-similarity in World Wide Web traffic—Evidence and possible causes, *in* "Proceedings of ACM Sigmetrics'96, 1996," pp. 160–169.
6. I. Daubechies, "Ten Lectures on Wavelets," SIAM, Philadelphia, 1992.
7. L. Delbeke and P. Abry, Wavelet based estimators for the self-similarity parameter of α-stable processes, preprint, 1998.
8. A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger, Dynamics of IP traffic: A study of the role of variability and the impact of control, *in* "ACM SIGCOMM Conf. Proc., 1999."
9. A. Feldmann, A. C. Gilbert, and W. Willinger, Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic, *in* "Proc. of the ACM/SIGCOMM'98, Vancouver, B.C., 1998," pp. 25–38.
10. A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, The changing nature of network traffic: Scaling phenomena, *Comput. Comm. Rev.* **28** (1998), 5–29.
11. R. A. Floyd, Short-term file reference patterns in a UNIX environment, Technical Report 177, Dept. of Computer Science, Univ. Rochester, 1986.
12. S. Jaffard, Exposants de Hölder en des points donnés et coéfficients d'ondelettes, *C. R. Acad. Sci. Paris* **308** (1989), 79–81.
13. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of ethernet traffic (extended version), *IEEE/ACM Trans. Networking* **2** (1994), 1–15.
14. B. Mandelbrot, A. Fisher, and L. Calvet, A multifractal model of asset returns, Technical Report, Cowles Foundation Discussion Paper 1164, Cowles Foundation, 1997.
15. Y. Meyer, "Wavlets, Vibrations and Scalings," CRM Monograph Series, Vol. 9, Amer. Math. Soc., Providence, RI, 1998.
16. J. K. Ousterhout, H. Da Costa, D. Harrison, J. A. Kunze, M. Kupfer, and J. G. Thompson, A trace-driven analysis of the UNIX 4.2BSD file system, Technical Report CSD-85-230, Dept. of Computer Science, Univ. California Berkeley, 1985.
17. S. Resnick, G. Samorodnitsky, A. Gilbert, and W. Willinger, Wavelet analysis of conservative cascade, preprint, 2000.
18. G. Samorodnitsky and M. Taqqu, "Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance," Chapman & Hall, London/New York, 1994.
19. S. Seuret and A. C. Gilbert, Pointwise Hölder exponent estimation in data network traffic, *in* "ITC Specialist Seminar, Monterey, CA, September 2000."